# Inference for Categorical Data

Shannon Pileggi

STAT 217

# OUTLINE

# Overview of Statistical Methods

Quantitative response (means)

- one sample t-test
- paired t-test
- two sample t-test
- anova

Categorical response (proportions)

- one sample z-test
- N/A for STAT 217
- two sample z-test
- chi-squared test

# Gardasil vaccinations

- HPV is a sexually transmitted virus with links to certain types of cancer
- the FDA approved the Gardasil vaccination in 2006 to protect against HPV
- Gardasil is a three shot sequence recommended for women age 9-26
- the study subjects are 1413 females aged 11-26 who
  1. made their first "Gardasil visit" to a Johns Hopkins Medical Institution clinic between 2006 and 2008, and
  2. had 12 months to complete the regimen

## Is this study observational or experimental?

1. observational
2. experimental

1. Describe the sample:

2. Describe the population:

# The first 6 observations in the data set

```
> head(gardasil)
  AgeGroup          Race Completed       InsuranceType
1    18-26         white       yes            military
2    18-26         white       yes            military
3    18-26         white        no       private payer
4    11-17         white       yes            military
5    11-17 other/unknown        no            military
6    11-17         black        no medical assistance
```

| | |
|---|---|
| AgeGroup | 11-17; 18-26 |
| Race | white, black, Hispanic, other/unknown |
| Completed | completion of three-shot sequence 12 months (yes, no) |
| InsuranceType | medical assistance, private payer, hospital based, military |

The Data

## One sample z-test

Two sample z-test

Chi-squared test

Summary

The research question: Suppose the CDC claims that the completion rate for the Gardasil vaccination is 35%. Do we have evidence for or against this claim?

```
> head(gardasil)
  Completed
1       yes
2       yes
3        no
4       yes
5        no
6        no

> addmargins(table(gardasil$Completed))

  no  yes  Sum
 944  469 1413
```

1. What types of variables do we have?
2. How many groups are we studying?
3. How can we approach the problem?

# Parameter of interest

## What is the parameter of interest?

1. whether or not a female completes the shot sequence
2. the observed sample proportion of females that complete the shot sequence
3. the population proportion of females that complete the shot sequence
4. the number of females that complete the shot sequence
5. the population mean number of shots that a female gets
6. if the observed sample proportion of females that complete the shot sequence is different than 0.35

# Answering the research question

We can answer the research question of interest with:

- A confidence interval for $p$
- A hypothesis test of $H_0$: $p = 0.35$ vs $H_a$: $p \neq 0.35$ (this is the one sample z-test)

The value of $p$ in our hypothesis statement is called $p_0$. Here, $p_0 = 0.35$. This is what we assume to be true about $p$ **under** the null hypothesis.

# The possibilities

469 out of 1413 (33.2%) completed the vaccination sequence

$H_0 : p = 0.35$ vs $H_a : p \neq 0.35$

1. It could be that the population proportion that completes the vaccination sequence really is 0.35, and we observed 33.2% by random chance in our sample. This idea corresponds to the null hypothesis.
2. It could be that the population proportion that completes the vaccination sequence is not 0.35. This idea corresponds to the alternative hypothesis.

# Conditions required for the one sample z-test

1. independent observations

2. at least 10 observed 'successes' and 10 observed 'failures'

# One sample z-test

The test statistic looks like a $z$-score $\left(z = \frac{x - \mu}{\sigma}\right)$:

$$\text{test statistic} = \frac{\text{sample proportion} - \text{null hypothesis proportion}}{\text{standard error under } H_0}$$

$$z = \frac{\hat{p} - p_0}{se_{p_0}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} =$$

This is a **z test statistic** because the test statistic follows a standard normal distribution (a normal distribution with a mean of 0 and a standard deviation of 1).

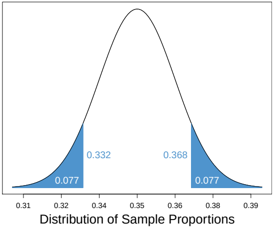
# Results in R

```
> prop.test(469,1413,p=0.35,correct=F)

        1-sample proportions test without continuity correction

data:  469 out of 1413, null probability 0.35
X-squared = 2.0308, df = 1, p-value = 0.1541
alternative hypothesis: true p is not equal to 0.35
95 percent confidence interval:
 0.3078495 0.3568977
sample estimates:
        p
0.3319179
```

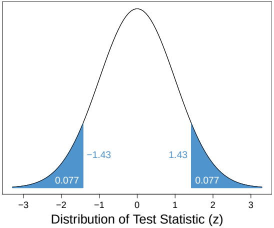
Even though we performed the one sample z-test, R reports a chi-squared test statistic. This has equivalent results to the $z$ test statistic because $z = \sqrt{\chi^2_1}$.

# Interpret the p-value

A sample proportion of $\hat{p} = 0.332$ is 1.43 standard errors below the claimed null hypothesis proportion of 0.35.




- If the proportion of females that complete the Gardasil vaccination sequence really is 0.35, the probability that we would see a sample proportion less than 0.332 or greater than 0.368 is 0.154.




- If the proportion of females that complete the Gardasil vaccination sequence really is 0.35, the probability that we would see a test statistic less than -1.43 or greater than 1.43 is 0.154.

# Conclusion in context

1. **Decision about $H_0$:**

2. **Statement about the parameter tested in context of the research question:**

3. **Provide a deeper connection of how this relates to the research question:**

**This decision could have been the result of... (mark all that apply)**

1. a Type I error
2. a Type II error
3. a correct decision

## A 95% confidence interval for the population proportion would...

1. include 0
2. not include 0
3. include 0.95
4. not include 0.95
5. include 0.35
6. not include 0.35
7. include 0.15
8. not include 0.15

# Review: Conditions required for the confidence interval

1. independent observations

   whether or not one female completes the shot sequence is independent from the other females' completion of the shot sequence

2. at least 10 'successes' and 10 'failures'

   we have 469 females that completed the sequence ($> 10$) and 944 females that did not complete the sequence ($> 10$) so this condition is satisfied

# Review: 95% confidence interval for $p$

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

based on a 95% conf. level
$z^* = 1.96$

$se = \sqrt{\frac{0.332\times(1-0.332)}{1413}} = 0.0125$

$0.332 \pm 1.96 \times 0.0125$

$0.332 \pm 0.025$

95% CI for $p$: $(0.307, 0.357)$

**What is the margin of error for this confidence interval?**

1. 0.0125
2. 0.025
3. 0.05
4. 0.95
5. 1.96

## Which of the following is a correct interpretation of the (0.31, 0.36) interval?

1. We are 95% confident that the population proportion of females that complete the Gardasil vaccination sequence is in the interval (0.31, 0.36).
2. We are 95% confident that among the 1413 females in this study, the proportion that complete the Gardasil vaccination sequence is in the interval (0.31, 0.36).
3. Between 31 to 36% of the time, we are 95% confident that we will reject the null hypothesis.
4. Between 31 to 36% of the time, we are 95% confident that we will find a statistically significant result.

The research question: Does completion rate differ by age group?

First six observations:

```
> head(gardasil)
  AgeGroup     Completed
1     18-26          yes
2     18-26          yes
3     18-26           no
4     11-17          yes
5     11-17           no
6     11-17           no
```

1. What types of variables do we have?
2. How many groups are we studying?
3. How can we approach the problem?

```
> addmargins(table(gardasil$AgeGroup,gardasil$Completed))
          no  yes  Sum
  11-17  454  247  701
  18-26  490  222  712
  Sum    944  469 1413
```

**Which two proportions should you compare to determine if completion rate differs by age group?**

1. 701/1413 vs 712/1413
2. 469/1413 vs 944/1413
3. 701/1413 vs 469/1413
4. 247/701 vs 222/712
5. 247/469 vs 222/469
6. 490/944 vs 222/469
7. 454/701 vs 247/701

Parameters of interest:

$p_1$ = population proportion of 11-17 year old females who complete the Gardasil vaccination sequence

$p_2$ = population proportion of 18-26 year old females who complete the Gardasil vaccination sequence

We can answer the research question of interest with:

- A hypothesis test of $H_0$: $p_1 = p_2$ vs $H_a$: $p_1 \neq p_2$ (this is the two sample z-test)
- A confidence interval for $p_1 - p_2$. Three possible scenarios:

| | | |
|---|---|---|
| No difference | $p_1 - p_2 = 0$ | $\rightarrow p_1 = p_2$ |
| Difference | $p_1 - p_2 > 0$ | $\rightarrow p_1 > p_2$ |
| Difference | $p_1 - p_2 < 0$ | $\rightarrow p_1 < p_2$ |

## Which of these scenarios represent a two sample comparison?

1. We take a random sample of 100 Cal Poly students and test proportion that participate in Greek life differs from 0.2.
2. We take a random sample of 100 Cal Poly students and test if the proportion that participate in Greek life differs among males and females.
3. We take a random sample of 100 Cal Poly students and ask them in their Sophomore and Senior year if they participate in Greek life, and we want to know if the proportion that participates in Greek life changes over year in school.
4. We take a random sample of 100 Cal Poly students and test if the proportion that participate in Greek life differs by where they grew up (ie, west, south, northeast, midwest).

# Conditions required for the CI for $p_1 - p_2$

1. independent observations

2. check to see that you have at least 10 observed ‘successes’ and 10 observed ‘failures’ in *each* group

# Confidence interval for $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times se$$

11-17 females: $\hat{p}_1 = 247/701 = 0.352$
18-26 females: $\hat{p}_1 = 222/712 = 0.312$

$$\begin{aligned} se &= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= \sqrt{\frac{0.352(1-0.352)}{701} + \frac{0.312(1-0.312)}{712}} \\ &= 0.025 \end{aligned}$$

# Interpret the CI

- We are 95% confident that the difference in the population proportion of 11-17 year old females and 18-26 year old females that complete the Gardasil vaccination sequence is in the interval (-0.009, 0.089).
- Because this interval includes 0, it is plausible that the population proportion of 11-17 year old females who complete the Gardasil vaccination sequence is equal to the population proportion of 18-26 year old females who complete the Gardasil vaccination sequence

Researchers want to know if wearing glasses differs by males and females. We calculated a 95% CI for the true proportion difference of men and women ($p_{men} - p_{women}$) who wear glasses who wear glasses to be (0.01,0.03). Which of the following statements is true? Mark all that apply.

1. In general, the proportion of men and women who wear glasses is small.
2. There is evidence that a higher proportion of males wear glasses compared to females.
3. There is evidence that a higher proportion of females wear glasses compared to males.
4. It is plausible that the proportion of females who wear glasses is equal to the proportion of males who wear glasses.

We calculated a 95% CI for the true proportion difference of men who wear glasses and women who wear glasses to be (0.01,0.03).

**If I used the same data to calculate a CI for $p_{women} - p_{men}$ instead of $p_{men} - p_{women}$, what would the CI be?**

1. the same
2. $(-0.03, 0.01)$
3. $(-0.03, -0.01)$
4. $(-0.01, 0.03)$
5. not enough information to determine

Consider doing the two sample z-test for $H_0$: $p_1 = p_2$ vs $H_a$: $p_1 \neq p_2$. Based on the 95% CI for $p_1 - p_2$ of (-0.009, 0.089), you would expect to

1. reject $H_0$
2. fail to reject $H_0$
3. accept $H_0$
4. find a statistically significant result
5. not enough information to determine

# Conditions required for the two sample z-test

1. independent observations

2. check to see that you have at least 10 observed 'successes' and 10 observed 'failures' in *each* group

# The test statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{se_0}$$

where $se_0$ is the standard error calculated assuming $H_0$ true.

1. $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{247 + 222}{701 + 712} = \dfrac{469}{1413} = 0.332$
2. 

$$\begin{aligned} se_0 &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} \\ &= \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\ &= \sqrt{0.332(1-0.332)\left(\frac{1}{701} + \frac{1}{712}\right)} \\ &= 0.025 \end{aligned}$$

# Results in R

```
> prop.test(c(247,222),c(701,712),correct=F)

        2-sample test for equality of proportions without continuity c

data:  c(247, 222) out of c(701, 712)
X-squared = 2.62, df = 1, p-value = 0.1055
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.008517967  0.089630022
sample estimates:
   prop 1    prop 2
0.3523538 0.3117978
```

Even though we performed the one sample z-test, R reports a chi-squared test statistic. This has equivalent results to the $z$ test statistic because $z = \sqrt{\chi_1^2}$.

$H_0$: $p_1 = p_2$ vs $H_a$: $p_1 \neq p_2$, $p$-value $= 0.1055$

**Which is a *correct* interpretation of this $p$-value?**

1. Fail to reject $H_0$.
2. The probability that the two proportions are equal is 0.1055.
3. The probability that we committed a type I error is 0.1055.
4. The probability that we observe a test statistic as extreme or more extreme than 1.60 if the two proportions really are equal is 0.1055.
5. The probability that we observe a test statistic as extreme or more extreme than 1.60 if the two proportions really are unequal is 0.1055.

# Conclusion in context

1. **Decision about $H_0$:**

2. **Statement about the parameter tested in context of the research question:**

3. **Provide a deeper connection of how this relates to the research question:**

Do people who drink caffeinated beverages have a higher occurrence of heart disease than people who do not drink caffeinated beverages? 200 caffeinated beverage drinkers and 150 non-caffeinated beverage drinkers report whether or not they have heart disease.

**To answer this question would you use proportions or means AND paired or not paired measurements?**

1. Two proportions from not paired measurements
2. Two proportions from paired measurements
3. Two means from not paired measurements
4. Two means from paired measurements

The Data

One sample z-test

Two sample z-test

## Chi-squared test

Summary

The research question: Does completion rate differ by insurance type?

First six observations:

```
> head(gardasil)
  Completed       InsuranceType
1       yes            military
2       yes            military
3        no       private payer
4       yes            military
5        no            military
6        no  medical assistance
```

1. What types of variables do we have?
2. How many groups are we studying?
3. How can we approach the problem?

```
> addmargins(table(gardasil$InsuranceType,gardasil$Completed))

                      no  yes  Sum
  hospital based      45   39   84
  medical assistance 220   55  275
  military           209  122  331
  private payer      470  253  723
  Sum                944  469 1413

> prop.table(table(gardasil$InsuranceType,gardasil$Completed),
      margin=1) #row proportion

                            no       yes
  hospital based     0.5357143 0.4642857
  medical assistance 0.8000000 0.2000000
  military           0.6314199 0.3685801
  private payer      0.6500692 0.3499308
```

# The chi-squared test

Equivalent ways of stating the hypotheses:

$H_0$: the two variables are independent

$H_a$: the two variables are dependent

$H_0$: the two variables are not associated

$H_a$: the two variables are associated

# Gardasil data

| | hospital based | medical assistance | military | private payer | Total |
|---|---|---|---|---|---|
| yes | | | | | 469 |
| no | | | | | 944 |
| Total | 84 | 275 | 331 | 723 | 1413 |

If there was no association between `Completed` and `InsuranceType` (ie, $H_0$ true), how many females would you expect to see complete the shot sequence on the hospital based insurance?

1. What is the overall proportion of completion?
2. Apply that to the total number on hospital based insurance.

# Expected cell counts

If $H_0$ true,

$$\text{Expected cell count} = \frac{\text{Row total} \times \text{Column total}}{\text{Total sample size}}$$

| | hospital based | medical assistance | military | private payer | Total |
|---|---|---|---|---|---|
| yes | 39 (27.9) | 55 (91.2) | 122 (109.9) | 253 (240.0) | 469 |
| no | 45 (56.1) | 220 (183.7) | 209 (221.1) | 470 (483.0) | 944 |
| Total | 84 | 275 | 331 | 723 | 1413 |

# Conditions for the $\chi^2$ test

1. observations are independent within each of the groups

2. expected cell count $\geq 5$ in all cells

# The test statistic

| | hospital based | medical assistance | military | private payer | Total |
|---|---|---|---|---|---|
| yes | 39 (27.9) | 55 (91.2) | 122 (109.9) | 253 (240.0) | 469 |
| no | 45 (56.1) | 220 (183.7) | 209 (221.1) | 470 (483.0) | 944 |
| Total | 84 | 275 | 331 | 723 | 1413 |

The **chi-squared test statistic** summarizes the difference between the observed and expected counts.

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

with $df = (\text{number of rows } -1) \times (\text{number of columns } -1)$

```
> chisq.test(gardasil$Completed,gardasil$InsuranceType,correct=F)

        Pearson's Chi-squared test

data:  gardasil$Completed and gardasil$InsuranceType
X-squared = 31.283, df = 3, p-value = 7.411e-07
```

**What can we conclude? ______________; we (do/do not) have any evidence of an association between completion and insurance type.**

1. Reject $H_0$; do
2. Reject $H_0$; do not
3. Fail to reject $H_0$; do
4. Fail to reject $H_0$; do not

If your data can go in a 2x2 contingency table, the two proportion $z$-test and the $\chi^2$ test will give you equivalent results.

```
> addmargins(table(gardasil$AgeGroup,gardasil$Completed))
         no  yes  Sum
  11-17  454  247  701
  18-26  490  222  712
  Sum    944  469 1413

> prop.test(c(247,222),c(701,712),correct=F)
> chisq.test(gardasil$AgeGroup,gardasil$Completed,correct=F)
```

Both test results yield a $p$-value of 0.1055.

**Two sample z-test:** Fail to reject $H_0$; we do not have evidence that the proportion of 11-17 year olds that complete the vaccination sequence differs from the proportion of 18-26 year olds that complete the vaccination sequence

**Chi-squared test:** Fail to reject $H_0$; there is no evidence of an association between age group and completion of the vaccination sequence

The Data

One sample z-test

Two sample z-test

Chi-squared test

Summary

# Attending religious services weekly by gender

| | **Case A** | | | **Case B** | | | **Case C** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Yes | No | n | Yes | No | n | Yes | No | n |
| Female | 52 | 48 | 100 | 104 | 96 | 200 | 5200 | 4800 | 10,000 |
| Male | 50 | 50 | 100 | 100 | 100 | 200 | 5000 | 5000 | 10,000 |
| | $\hat{p}_F = 0.52$<br>$\hat{p}_M = 0.50$<br>$\chi^2 = 0.08$<br>$p$-value$= 0.78$ | | | $\hat{p}_F = 0.52$<br>$\hat{p}_M = 0.50$<br>$\chi^2 = 0.16$<br>$p$-value$= 0.69$ | | | $\hat{p}_F = 0.52$<br>$\hat{p}_M = 0.50$<br>$\chi^2 = 8.0$<br>$p$-value$= 0.005$ | | |

Case (A/B/C) shows the strongest association between gender and attendance; case (A/B/C) shows the strongest evidence of an association between gender and attendance.

# Interpreting the *p*-value

The *p*-value represents the **strength** of the **evidence**:

- small *p*-values mean you have strong evidence of an association between two variables
- small *p*-values do not mean you have evidence of a strong association between two variables
- large *p*-values mean there is no evidence of an association

Other measures represent the **strength** of the **association**:

- difference of means: $(\bar{x}_1 - \bar{x}_2)$
- difference of proportions: $(\hat{p}_1 - \hat{p}_2)$

The **strength** of the **association** can help you assess if an observed difference is meaningful.

# Group Exercise

A researcher investigated if the proportion of non-athletes (group 1) and athletes (group 2) that are on the Dean's List differs.

```
data:  c(1000, 1100) out of c(10000, 10000)
X-squared = 5.3206, df = 1, p-value = 0.02108
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.018495941 -0.001504059
sample estimates:
prop 1 prop 2
  0.10   0.11
```

What is the best conclusion from this analysis? We (do/do not) have strong statistically significant evidence that the proportion of students on the Dean's list differs by athlete status, and the effect of athlete status is (strong/weak).

1. do; strong
2. do not; strong
3. do; weak
4. do not; weak

# Different methods

| Method | Use | Variables | Estimation | Testing |
|---|---|---|---|---|
| Single mean *(one-sample t-test)* | quantitative response in single group | one quantitative variable | CI for $\mu$ | $H_0$: $\mu = \mu_0$ |
| *Two means *(two-sample t-test)* | quantitative response in two groups | one quantitative variable and one categorical variable | CI for $\mu_1 - \mu_2$ | $H_0$: $\mu_1 = \mu_2$ |
| Dependent means *(paired t-test)* | quantitative response measured on same observation | two paired quantitative variables | CI for $\mu_d$ | $H_0$: $\mu_d = 0$ |
| *ANOVA | quantitative response in $> 2$ groups | one quantitative variable and one categorical variable | Tukey pairwise intervals | $H_0$: $\mu_1 = \mu_2 = \cdots = \mu_g$ |
| Single proportion *(one-sample z-test)* | categorical response in single group | one categorical variable | CI for $p$ | $H_0$: $p = p_0$ |
| *Two proportions *(two-sample z-test)* | categorical response in two groups | two categorical variables | CI for $p_1 - p_2$ | $H_0$: $p_1 = p_2$ |
| *$X^2$ test | categorical response in $\geq 2$ groups | two categorical variables | N/A | $H_0$: no association/ vars independent |

*The starred methods can answer the question "Is there an association?" If we reject $H_0$, then we conclude that some sort of association is present in the two variables.